We've Lost Control of AI

Caption: A clip from a video by SciShow that explains what sandbagging is.

We say that an AI is "sandbagging" if it *intentionally* underperforms on an evaluation. This could happen if it is being "lazy" in some way, or if it attempts to hide its capabilities.[1] The term comes from "sandbagging" in sports or games like poker meaning underplaying one's strength, which in turn is based on sneaking up on someone and hitting them with a bag of sand.

Another kind of "sandbagging" can happen when the *developers* of a model intentionally fail to elicit its full capabilities in order to, e.g., avoid reaching a threshold that would trigger stricter legal requirements.

Alternative phrasings

-

Related

-

Scratchpad

-

[1] Jan Leike considers that the latter is not sandbagging.